Manual Comparison analysis

Author: Kaiden R. Sewradj
Date of last update: 28/03/2025

Supervised by: Quentin Helleu, Ellen Carbo & Jean-Paul Concordet

0. Table of Contents

0. Table of Contents....2

1. Requirements.... 3

2. Usage.... 4

2.1. Clone repo.... 4

2.1.1. What is in the repo?.... 4

2.2. Set up config file.... 4

2.4. Run pipeline.... 5

2.4.1. Running for the first time.... 5

2.4.2. Running with new data....5

3. Output.... 5

3.1. logfiles.... 5

3.2. sonicOut.... 5

4. Get a summary.... 5

5. References.... 7

1. Requirements

This pipeline was designed to work with the Slurm workload manager and relies on this system to dispatch jobs. Other requirements are:

- R v4.4.1 (R Core Team, 2023)
    - clusterProfiler v4.12.6 (Yu, 2024)
    - dplyr v1.1.4 (Wickham *et al.*, 2023)
    - enrichplot v1.24.4 (Yu, 2025)
    - ggplot2 v3.5.1 (Wickham, 2016)
    - stringr v1.5.1 (Wickham, 2023)
    - tidyr v1.2.1 (Wickham *et al.*, 2024)
    - ontologyIndex v2.12 (Greene *et al.*, 2016)

R is loaded with Modules v4.6.1 as follows: `module load r`. All R packages are expected to be pre-installed.

2. Usage

2.1. Clone repo

```
git clone https://github.com/Kaiden-exe/comparison_analysis
```

2.1.1. What is in the repo?

- `scripts` [directory]
  - All scripts are here
- `config.sh`
  - Template configuration file.
- `comparison_analysis.sh`
  - Main script file that dispatches the Slurm jobs
- Manual Comparison Analysis.pdf
  - You are here.

2.1. Prepare files

- `sonicOutput=ortholog_groups.tsv`
  - Orthology clusters in the format as generated by SonicParanoid (Consentino *et al.*, 2024)
- `species=species.txt`
  - This file is a list of speciesIDs separated by newlines. It is assumed that the columns are speciesID.fasta
- `GO_OBO=go.obo`
  - You can download the go.obo file here.

2.2. Set up config file

- `DEGout=DEG_out`
    - Directory with DEG tables. The tables are expected to be named speciesID_DESeq_results.txt with the speciesID corresponding to the previously mentioned species.txt
- `transmaps=transmaps`
    - Directory with tsv files that link geneID (first column) to proteinID (second column). These tables are expected to be named speciesID.gene_trans_map with the speciesID corresponding to the previously mentioned species.txt.
    - If you ran the transcriptomics analysis pipeline there is a script included to create these files. Run:
      `bash scripts/trin_trans_map.sh <transdecoderOut> transmaps`
- `emapperOut=emapper_out`
    - Directory with speciesID.emapper.annotations files. These files are the output of eggnog-mapper (Cantalapiedra *et al*., 2021). The speciesID must correspond to the previously mentioned species.txt

2.4. Run pipeline

```
bash comparison_analysis.sh -c config.sh
```

3. Output

3.1. logfiles

All stdout and stderror logfiles are in here. Filenames include batch number and array number if applicable.

3.2. DEG_OG

- `orthologgroups_converted.tsv`
    - ortholog_groups.tsv reformatted as output of OrthoFinder (Emms & Kelly, 2019)
- `speciesID_DEG_OG.tsv`
    - DEG tables with the corresponding OG(s) that gene is in

3.3. GO_enrichment

- TERM2GENE.tsv & TERM2NAME.tsv
    - Files linking GO terms to genes or their description
- dotplot.png
    - Enrichment of significantly differentially expressed DEGs per species
- clusterProfiler.RData
    - Enrichment analysis image

5. References

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution,* 38(12), 5825–5829. DOI: 10.1093/molbev/msab293

Cosentino, S., Sriswasdi, S. & Iwasaki, W. SonicParanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. *Genome Biol,* **25**, 195 (2024). DOI: 10.1186/s13059-024-03298-4

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**(1). DOI: 10.1186/s13059-019-1832-y

Greene, D., Richardson, S., & Turro, E. (2016). ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*, **33**(7), 1104–1106. DOI: 10.1093/bioinformatics/btw763

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Yu, G. (2024). Thirteen years of clusterProfiler. *The Innovation*, **5**(6), 100722. DOI: 10.1016/j.xinn.2024.100722

Yu, G. (2025). enrichplot: Visualization of Functional Enrichment Result. R package version 1.26.6, https://yulab-smu.top/biomedical-knowledge-mining-book/

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H. (2023). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.1, https://github.com/tidyverse/stringr, https://stringr.tidyverse.org.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org

Wickham H, Vaughan D, Girlich M (2024). tidyr: Tidy Messy Data. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org